

大数据基础知识

大数据并行计算

什么是大数据

- 大数据是传统数据处理软件**难以处理**的**复杂**数据集；
- 大数据的核心是数据**存储与管理**、数据**处理与分析**。

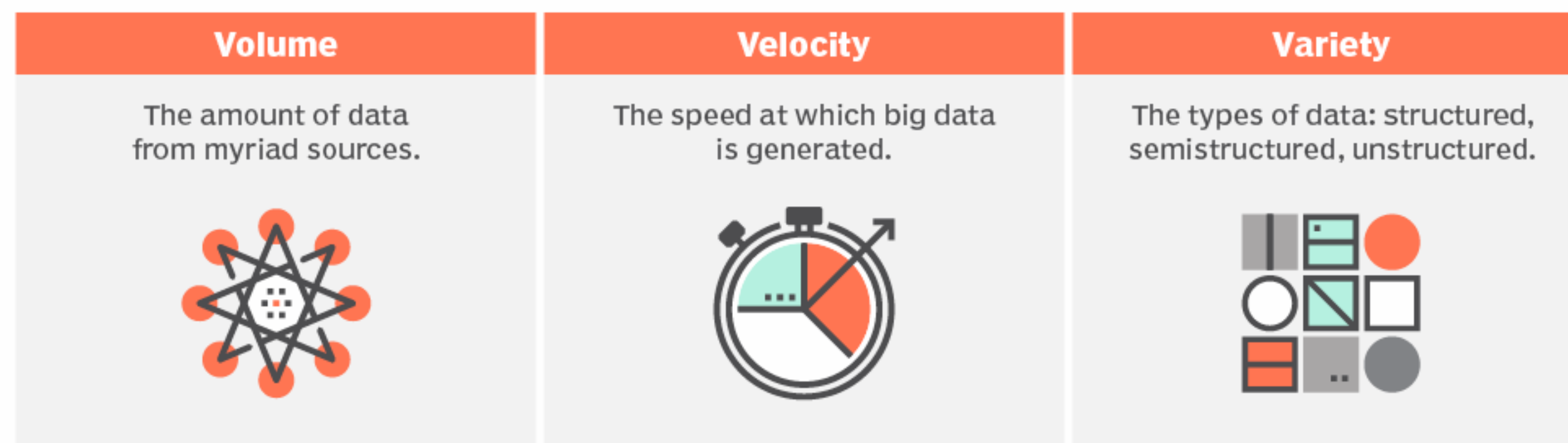


大数据的3个V




- 体量（volume），即数据的大小；
- 多样性（variety），即不同的来源和格式；
- 速度（velocity），即数据的速度。

The 3 V's of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3 V's: *volume*, *velocity* and *variety*.



大数据处理系统

Spark  vs  Hadoop MapReduce		
Factors	Spark 	Hadoop MapReduce
Speed	100x times than MapReduce	Faster than traditional system
Written In	Scala	Java
Data Processing	Batch / real-time / iterative / interactive / graph	Batch processing
Ease of Use	Compact & easier than Hadoop	Complex & lengthy
Caching	Caches the data in-memory & enhances the system performance	Doesn't support caching of data

MapReduce编程框架

- MapReduce是一种编程模型；
- 并行计算过程抽象为Map和Reduce两个函数；
- MapReduce的核心是“分而治之”的策略；
- 数据经历输入-切分-转换-洗牌-合并-输出的过程；
- Master/Slave架构，Master用于调度，Slave执行：
 - Map任务（切分和转换）；
 - Reduce任务（洗牌和合并）。

在阿里云安装配置Hadoop

- 远程登录云服务器
 - 获取IP地址，用户名和密码

☰

阿里云

工作台

搜索...

云服务器 ECS

概览

事件

标签

自助问题排查

应用管理

我的常用

实例与镜像

实例

镜像

概览

资源搜索

资源报表

ECS使用成熟度评估与洞察

我的资源

云服务器

运行中

即将过期

1

1

0

创建实例

迁移上云

可按ID、名称、IP等属性模糊搜索云服

i-0jlgktd9g7z03u7j1kly

运行中 (2核(vCPU) 2GiB)

名称

iZ0jlgktd9g7z03u7j1klyZ

地域

华北6 (乌兰察布)

创建时间

2023年8月31日 20:29:00

公网IP

8.130.34.86

在阿里云安装配置Hadoop

- 远程登录云服务器
 - 获取IP地址，用户名和密码

云服务器 ECS

概览

事件

标签

自助问题排查

应用管理

我的常用

实例与镜像

实例

云服务器 ECS / 实例 / i-0jlgktd9g7z03u7j1kly

← iZ0jlgktd9g7z03u7j1klyZ ▾

实例详情

监控

安全组

云盘

快照一致性组

快照

弹性网卡

定时与自动化任务

操作记录

健康诊

基本信息

实例问题排查 NEW | 启动 | 重启 | 停止 | 配置安全组规则 | 重置实例密码 | ⋮

iZ0jlgktd9g7z03u7j1klyZ

🔗

✔ 运行中

实例ID

i-0jlgktd9g7z03u7j1kly

远程连接

地域

华北 6 (乌兰察布)

资源组

-

公网IP

8.130.34.86

公网IP转换为弹性公网IP

在阿里云安装配置Hadoop

- 远程登录云服务器

- ssh远程登录

- Mac系统：打开终端；



- Windows系统：搜索终端，并打开；

- 输入 `ssh 用户名@IP地址`；

- 输入密码；

- SSH (Secure Shell, 安全外壳) 是一种网络安全协议，通过加密和认证机制实现安全的访问和文件传输等业务。

```
ben — ssh root@8.130.34.86 — 80x24
Last login: Thu Sep 14 22:26:50 on ttys001
[ben@jiumingwobianchengMacBookPro14cunledeMacBook-Pro ~ % ssh root@8.130.34.86 ]
[root@8.130.34.86's password: ]
Permission denied, please try again.
[root@8.130.34.86's password: ]

Welcome to Alibaba Cloud Elastic Compute Service !

Updates Information Summary: available
    46 Security notice(s)
        25 Important Security notice(s)
        21 Moderate Security notice(s)
Run "dnf upgrade-minimal --security" to apply all updates. More details please refer to:
https://help.aliyun.com/document_detail/416274.html
Last failed login: Thu Sep 14 22:41:44 CST 2023 from 223.72.88.191 on ssh:notty
There were 2 failed login attempts since the last successful login.
Last login: Sun May 29 15:42:16 2022 from 101.86.101.255
[root@iZ0jlgktd9g7z03u7j1klyZ ~]#
```


在阿里云安装配置Hadoop

- 安装Docker
 - Docker是容器化技术；
 - Docker对进程进行封装隔离，属于操作系统层面的虚拟化技术。由于隔离的进程独立于宿主和其它的隔离的进程，因此也称其为容器。
 - 借助Docker，可以将容器当作重量轻、模块化的虚拟机来使用；
 - 通过构建不同容器充当不同实体机，可实现分布式集群的搭建。

在阿里云安装配置Hadoop

- 更新yum包
 - yum (Yellow dog Updater, Modified) 是一个在 Fedora 和 RedHat 以及 SUSE 中的 Shell 前端软件包管理器。基于 RPM 包管理, 能够从指定的服务器自动下载 RPM 包并且安装, 可以自动处理依赖性关系, 并且一次安装所有依赖的软件包, 无须繁琐地一次次下载、安装。
 - yum 提供了查找、安装、删除某一个、一组甚至全部软件包的命令, 而且命令简洁而又好记。
- 输入 `sudo yum update`
- 【注意: `sudo`是以管理员权限执行命令, 随意使用可能造成严重的后果, 一定要知道自己在做什么! 】

在阿里云安装配置Hadoop

- 安装需要的软件包
 - 输入 `sudo yum install -y yum-utils device-mapper-persistent-data lvm2`
 - `yum install`是安装软件包的命令；
 - `-y`表示安装过程选项全部选yes；
 - `yum-utils`是一系列yum工具的集合；
 - `device-mapper-persistent-data`和`lvm2`都是Device Mapper需要的软件包。

在阿里云安装配置Hadoop

- 设置yum源
 - 输入 `sudo yum-config-manager --add-repo https://download.docker.com/linux/centos/docker-ce.repo`
- 安装最新版的Docker CE
 - 输入 `sudo yum install docker-ce`
 - CE(Community Edition)是社区版，用于为了开发人员或小团队创建基于容器的应用。

在阿里云安装配置Hadoop

- 启动Docker
 - 输入 `systemctl start docker`
- 通过运行hello-world镜像验证Docker CE已被正确安装
 - 输入 `sudo docker run hello-world`

```
Hello from Docker!
```

```
This message shows that your installation appears to be working correctly.
```

在阿里云安装配置Hadoop

- 拉取Hadoop镜像
 - 输入 `docker pull registry.cn-beijing.aliyuncs.com/bitnp/docker-spark-hadoop`
- 使用docker images查看是否下载成功
 - 输入 `docker images`
 - docker images可以列出本地镜像。

REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
hello-world	latest	9c7a54a9a43c	4 months ago	13.3kB
registry.cn-beijing.aliyuncs.com/bitnp/docker-spark-hadoop	latest	8b768e1604ad	5 years ago	2.11GB

在阿里云安装配置Hadoop

- 在Hadoop镜像里创建三个容器（Master, Slave1, Slave2）
- 输入 `docker run -it --name Master -h Master registry.cn-beijing.aliyuncs.com/bitnp/docker-spark-hadoop /bin/bash`
 - `-i` 以交互模式运行容器，通常与 `-t` 同时使用；
 - `-t` 为容器重新分配一个伪输入终端，通常与 `-i` 同时使用；
 - `--name` 为容器指定一个名称；
 - `-h` 指定容器的hostname（主机名称/节点名称）；
 - [Read more.](#)

在阿里云安装配置Hadoop

- Master空容器已创建。
 - 现在我们已经是在容器里了； `[root@Master local]#`
 - Ctrl+P+Q返回初始目录，但不退出Master容器；
 - Ctrl+C返回初始目录，且退出Master容器。
- 依次创建Slave1和Slave2容器
 - 输入`docker run -it --name Slave1 -h Slave1 registry.cn-beijing.aliyuncs.com/bitnp/docker-spark-hadoop /bin/bash`
 - 输入`docker run -it --name Slave2 -h Slave2 registry.cn-beijing.aliyuncs.com/bitnp/docker-spark-hadoop /bin/bash`

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
 - 输入 `docker attach Master` 进入Master容器环境；
 - 在Master环境中下载vim、openssh-clients和openssh-server
 - 输入 `yum -y install vim openssh-clients openssh-server`
 - vim是一个文本编辑器；
 - Openssh是SSH透过计算机网络加密通信的实现。

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
- 配置Master容器的ssh密钥，依次输入
 - `/usr/sbin/sshd`
 - `/usr/sbin/sshd-keygen -A`
 - `/usr/sbin/sshd`
 - `ssh-keygen -t rsa`

```
Your identification has been saved in /root/.ssh/id_rsa.  
Your public key has been saved in /root/.ssh/id_rsa.pub.  
The key fingerprint is:  
SHA256:nx293z1qxrGIKC5BURRV300Cx98UbzVDh+NiPrCd1lA root@Master  
The key's randomart image is:  
+---[RSA 2048]---+  
|    o+o... .oo.B+|  
|    .      . o.*EB|  
|    .      . oo++|  
|    .      oo.o.|  
|    . S  o..o .|  
|    .      . 000.=|  
|    .      . + +o=o.|  
|    . . . . . =.o+|  
|    o..      o.. =|  
+-----[SHA256]-----+  
[root@Master local]#
```

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
 - 将密钥存入指定文件夹
 - 输入 `cat /root/.ssh/id_rsa.pub >> /root/.ssh/authorized_keys`
- 配置相应文件
 - 输入 `vim /etc/ssh/sshd_config`
 - 输入i进入编辑模式，找到Port 22位置并修改

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
 - 配置相应文件

#Port 22

#PermitRootLogin yes

#PubkeyAuthentication yes

#PasswordAuthentication yes

#ChallengeResponseAuthentication no

#UsePAM yes

#PrintLastLog no

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
- 配置相应文件
 - 按下Esc进入命令模式，输入wq（保存退出）；
 - 输入 `vim /etc/ssh/ssh_config`；
 - 找到StrictHostKeyChecking ask，将#去掉并把ask 改为no，并保存；
 - Ctrl+P+Q返回初始目录。

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
- 查看三个容器的ip地址
- 输入 `docker inspect -f '{{.Name}} - {{.NetworkSettings.IPAddress}}'`
`Master Slave1 Slave2`;

```
/Master - 172.17.0.2  
/Slave1 - 172.17.0.3  
/Slave2 - 172.17.0.4
```

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
- 进入Master容器[docker attach Master](#)
- 输入 [vim /etc/hosts](#);
- 将Slave1和Slave2及其对应的IP地址填上。

```
127.0.0.1      localhost
::1           localhost ip6-localhost ip6-loopback
fe00::0       ip6-localnet
ff00::0       ip6-mcastprefix
ff02::1       ip6-allnodes
ff02::2       ip6-allrouters
172.17.0.2     Master
172.17.0.3     Slave1
172.17.0.4     Slave2
```

在阿里云安装配置Hadoop

- 使用ssh把三个容器连接起来。
 - 退出Master容器，分别进入Slave1和Slave2重复上述配置；
 - 将三个容器的密钥同时放在每个容器的/root/.ssh/authorized_keys文件中；
 - 在Master容器中输入ssh Slave1和ssh Slave2进入Slave1和Slave2；
 - 输入logout退出。

在阿里云安装配置Hadoop

- 如果因为某些原因容器关闭了（以Master容器为例），例如服务器重启，或者输入了 `docker stop Master`。
- 重启容器 `docker start Master`；
- 进入容器 `docker attach Master`；
- 重新配置 `vim /etc/hosts`；
- 重启sshd `/usr/sbin/sshd`；

在阿里云安装配置Hadoop

- 依次配置每个容器的core-site.xml、yarn-site.xml、mapred-site.xml以及pdfs-site.xml文件。
- 输入`find / -name core-site.xml`查找路径；
- 输入`find / -name yarn-site.xml`查找路径；
- 输入`find / -name mapred-site.xml`查找路径；
- 输入`find / -name hdfs-site.xml`查找路径；
- 使用`vim`进入文件修改配置，注意配置`hdfs-site.xml`时Master和Slave不同。

在阿里云安装配置Hadoop

- core-site.xml配置

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://Master:9000</value>
  </property>
  <property>
    <name>io.file.buffer.size</name>
    <value>131072</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/usr/local/hadoop-2.7.5/tmp</value>
  </property>
</configuration>
```

在阿里云安装配置Hadoop

- yarn-site.xml配置

```
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>Master:8032</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>Master:8030</value>
  </property>
  <property>
    <name>yarn.resourcemanager.resource-tracker.address</name>
    <value>Master:8031</value>
  </property>
  <property>
    <name>yarn.resourcemanager.admin.address</name>
    <value>Master:8033</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>Master:8088</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

在阿里云安装配置Hadoop

- mapred-site.xml配置

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```


在阿里云安装配置Hadoop

- [hdfs-site.xml](#)配置
- Master容器

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/usr/local/hadoop-2.7.5/hdfs/name</value>
  </property>
</configuration>
```

- Slave1和Slave2容器

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.datanode.name.dir</name>
    <value>file:/usr/local/hadoop-2.7.5/hdfs/name</value>
  </property>
</configuration>
```

在阿里云安装配置Hadoop

- 在Master容器通过ssh连接Slave1和Slave2，删除其hdfs所有目录并重新创建。
 - 输入ssh Slave1
 - 输入rm -rf /usr/local/hadoop-2.7.5/hdfs
 - 输入mkdir -p /usr/local/hadoop-2.7.5/hdfs/data
 - 输入logout返回Master容器；
 - rm是移除命令，mkdir创建目录。

在阿里云安装配置Hadoop

- 在Master容器删除其hdfs所有目录并重新创建。
 - 输入`rm -rf /usr/local/hadoop-2.7.5/hdfs`
 - 输入`mkdir -p /usr/local/hadoop-2.7.5/hdfs/name`
 - 格式化NameNode HDFS目录
 - 输入`hdfs namenode -format`
- 在Master容器输入`vim /usr/local/hadoop-2.7.5/etc/hadoop/slaves`并修改。

在阿里云安装配置Hadoop

- 进入sbin文件，启动Hadoop集群
 - 输入`cd /usr/local/hadoop-2.7.5/sbin;`
 - 输入`./start-all.sh;`
 - （然后阿里云就炸了
 - 输入`jps`查看namenode是否启动。
- 输入`vim /etc/profile`配置文件，在末尾加上：

```
export JAVA_HOME=/usr/local/jdk1.0.0_162
export HADOOP_HOME=/usr/local/hadoop-2.7.5
export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

在阿里云安装配置Hadoop

- 输入 `source /etc/profile`;
- 进入Slave1容器并使用jps查看datanode是否启动;
- 回到Master容器中查看各容器启动状态
 - 输入 `hadoop dfsadmin -report`

在阿里云安装配置Spark

- 安装Anaconda。
 - 输入 `wget https://mirrors.aliyun.com/anaconda/archive/Anaconda3-5.1.0-Linux-x86_64.sh`;
 - 输入 `sudo bash Anaconda3-5.1.0-Linux-x86_64.sh`;
 - 更换安装路径至 `/usr/local/anaconda3`;
 - 将Anaconda添加至环境变量，输入`sudo vim /etc/profile` 并在末尾添加：

```
export PATH=/usr/local/anaconda3:$PATH
```


在阿里云安装配置Spark

- 安装Anaconda。
 - 激活环境变量：
 - 输入 `source /etc/profile`;
 - 输入 `source ~/.bashrc`。
- 安装JAVA，依次输入：
 - `sudo yum install -y bzip2`
 - `sudo yum search java|grep jdk`
 - `sudo yum install java-1.8.0-openjdk`（版本根据第2步列出来的有不同

在阿里云安装配置Spark

- 安装Spark。
 - 输入 `wget https://mirrors.tuna.tsinghua.edu.cn/apache/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz`;
 - 创建目录 `sudo mkdir -p /usr/local/spark`;
 - 复制文件 `sudo cp -r spark-3.5.0-bin-hadoop3.tgz /usr/local/spark`;
 - `cd /usr/local/spark`;
 - `sudo tar -zxvf spark-3.5.0-bin-hadoop3.tgz`。

在阿里云安装配置Spark

- 安装Spark。
 - 输入 `vim ~/.bash_profile` 在末尾添加：

```
export SPARK_HOME=/usr/local/spark/spark-3.5.0-bin-hadoop3
export PATH=$PATH:$SPARK_HOME/bin
```

- 激活配置文件 `source ~/.bash_profile`；
- 验证是否安装成功 `spark-shell`。

在阿里云安装配置Spark

- 安装Spark。
 - 输入 `:quit` 返回原目录；
 - 修改配置：
 - `cd /usr/local/spark/spark-3.5.0-bin-hadoop3/conf`
 - `sudo cp log4j2.properties.template log4j2.properties`
 - `sudo vim log4j2.properties`

在阿里云安装配置Spark

- 安装PySpark。
 - `pip install pyspark`;
 - `pip install findspark`。

作业

- 尝试在阿里云上安装Spark (P39-42)
- 了解Linux常用命令 (P43-44)